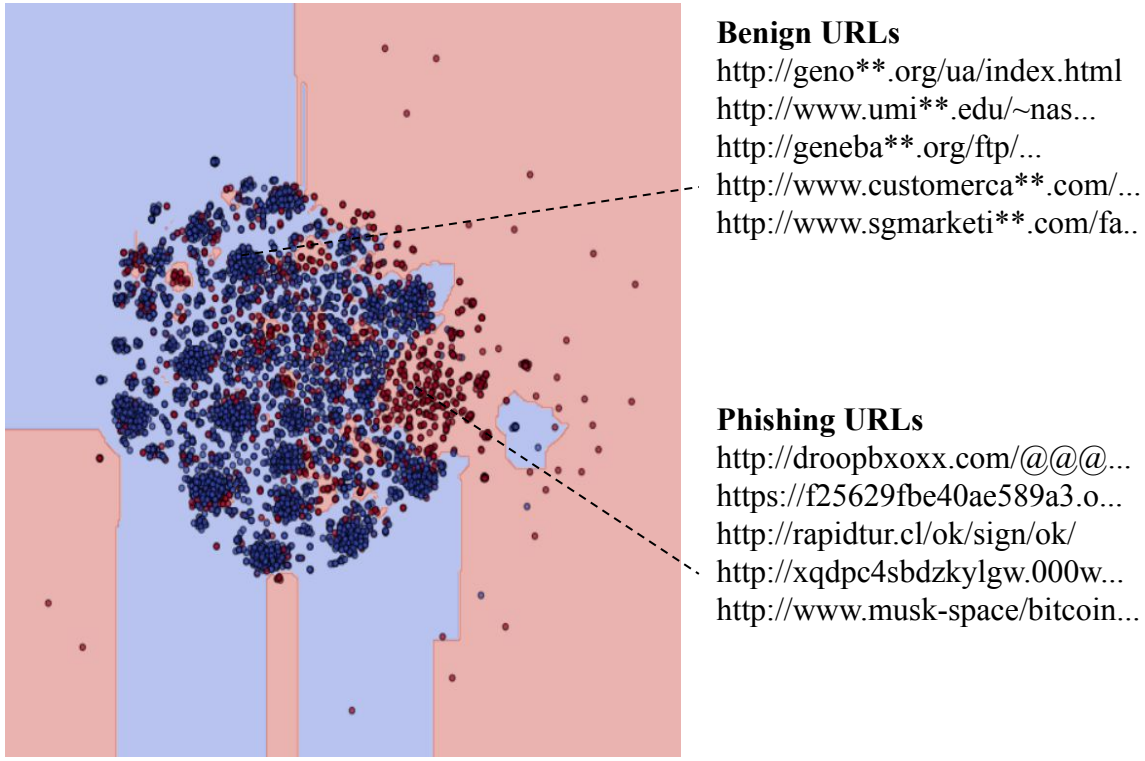
Zero-day 피싱 웹사이트 탐지를 위한 컨볼루션 오토인코더 기반 문자수준 URL 모형

Deep Character-level URL Model based on Convolutional Autoencoder for Zero-day Phishing Website Detection

피싱 웹사이트 URL 분류 문제

- 피싱/정상 URL 분류



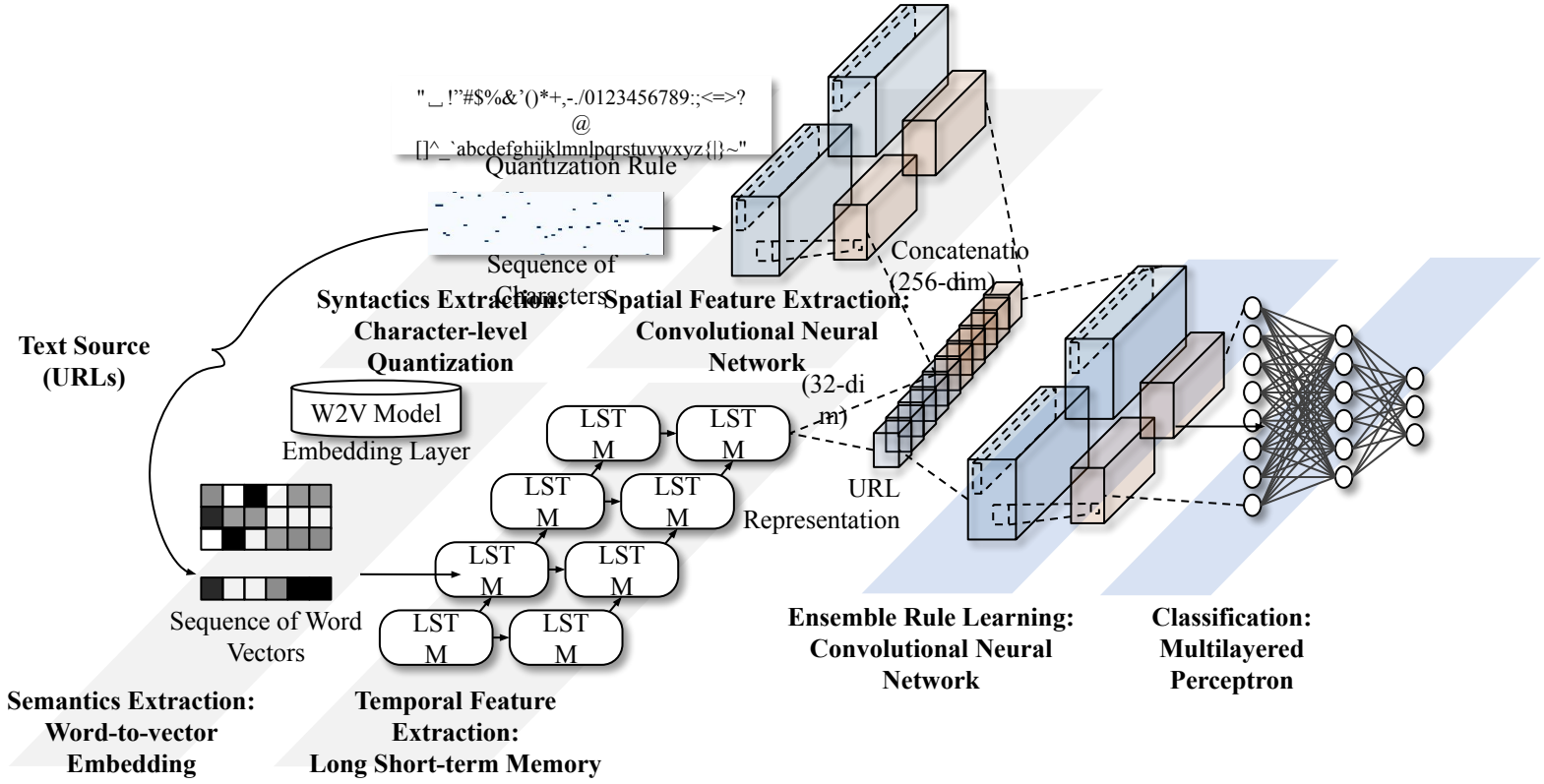
- 딥러닝 분류기 한계: 피싱 공격 특성
  - Data Imbalance (Yang et al. 2018.)
  - Zero-day Attack (Anand et al. 2019.)

기존 기계학습 기반 피싱 URL 분류 시도

Author	Feature Extraction	Feature Modeling	Description
2009	BOW	Naive Bayes	기계학습기반 URL분류 타당성검증을 위한 확률적 맵핑
2010	Lexical Features	Matching Rules	피씽 URL의 특징을 모델링할 수 있는 규칙설계
2010	BOW	SVM	대표적인 언어 모델링을 위한 전처리와 기계학습 방법 기반 피씽URL 특징 모델링
2015	Lexical Features	Random Forest	피씽 URL의 비선형적 특징벡터의 모델링을 위한 기계학습 모형의 앙상블
2017	W2V	LSTM	게이트 순환신경망 기반 단어벡터의 시퀀스 모델링
2018	W2V	GRU	
2018	Lexical Features	GAN	GAN 모형에 기반하여 URL 특징의 의미공간을 생성하고 가상의 피씽 URL을 생성, 모델링
2019	W2V	CNN-LSTM	CNN, LSTM을 직렬적으로 연결하여 단어벡터로부터 로컬한 공간적 특징과 글로벌한 시계열 모델링

기존 딥러닝 기반 피싱URL 탐지 시도

- URL: 구문적(문자의 시퀀스) 및 의미적(단어의 시퀀스) 특징을 내재
- 수준별 특징 모델링을 위한 딥러닝 융합 시도 (김혜정, 2019.)
  - 문자의 시퀀스: 컨볼루션 신경망(CNN) 으로 모델링
  - 단어의 시퀀스: 임베딩 후 순환 신경망(LSTM) 으로 모델링
  - 두 신경망의 융합: CNN 으로 앙상블 규칙 학습



기존 한계 대비 제안하는 방법

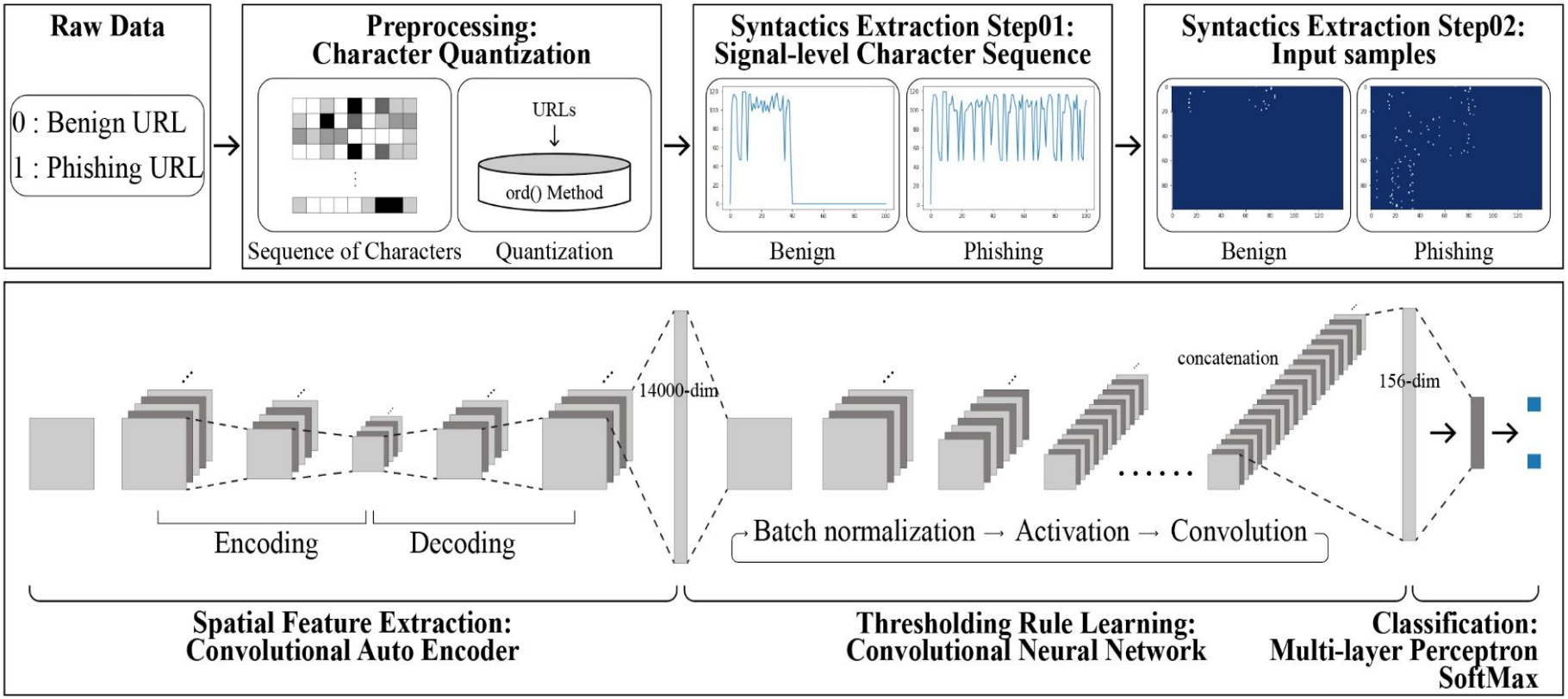
- 기존 방식의 한계 : 결정적 모형
  - 피싱 점수가 유클리디안 거리 기반
  - 원본 데이터 확률적 모델링 필요 (e.g. KL)

- 해결 1: 컨볼루션 오토인코더(CAE) 기반 정상 URL 모형 파라미터 θ* 모델링
  - $\theta^* = argmin_\theta \sum_{x_i \in X_train} L_D(\theta; x_i) = argmin_\theta \sum_{x_i \in X_train} ||f_\theta(x_i) - x_i||^2$
  - D : 학습 데이터의 클래스 분포
  - $f_\theta : X_{train} \rightarrow R$
  - 정상 URL의 재구축 오류 << 피싱 URL의 재구축 오류

- 해결 2: 재구축 오류의 Thresholding 값의 딥러닝 기반 학습

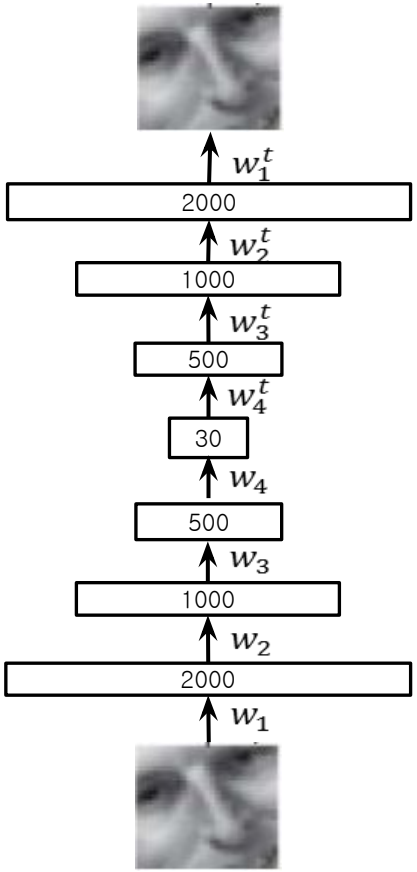
제안하는 방법 오버뷰

- CAE(Convolutional AutoEncoder)와 딥러닝 네트워크(CNN + MLP)의 결합 구조
- CAE-Thresholding 한계
  - CAE의 출력값을 1차원 경계로 분류하는 방법으로 문제 해결이 어려움
  - 제안: 연결된 신경망에 기반한 Thresholding 값의 비선형적 학습



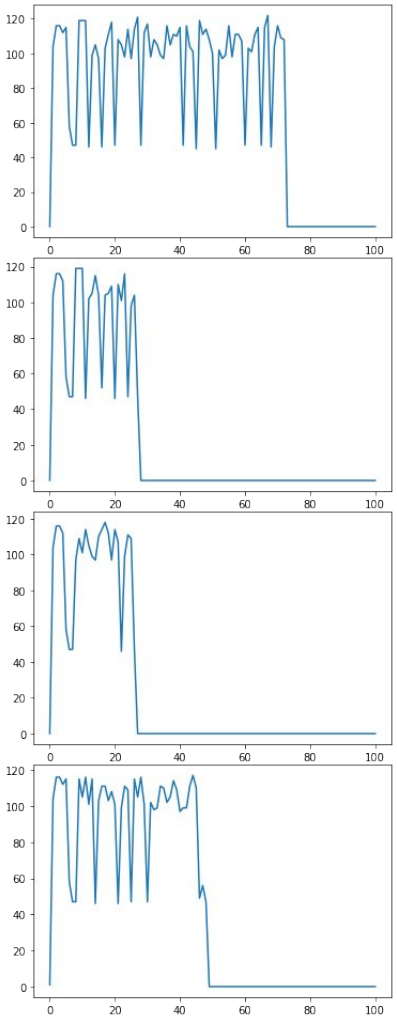
핵심 아이디어: CAE기반 정상 URL 모형

- Convolutional Auto Encoder
  - Encoding: 낮은 차원 특징벡터로의 사상
  - Decoding: 특징벡터로부터의 원본 데이터 재구축
  - High dimensional data Low dimensional code
- Thresholding CNN
- 논문의 기여
  - 정상 URL 의 잠재변수 (압축된 벡터 z) 모델링
  - 피싱 URL 재구축 오류 기반 피싱공격 탐지
    - Zero-day Attack 대처
    - Data Imbalance 문제 해결
  - 재구축 오류의 Thresholding값 학습: 비선형 경계
- 추후 연구 이슈: 효율적인 초기 weight 최적화 방법



데이터 전처리

Step01:Signal-level Character Sequence



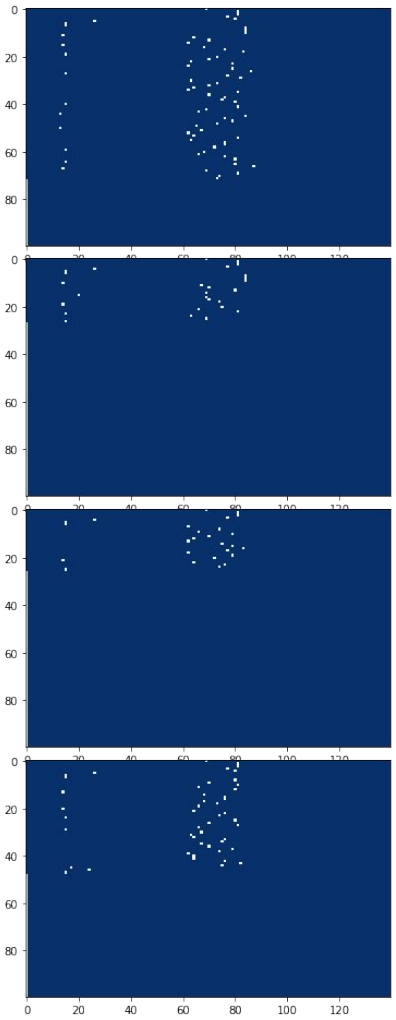
Phishing

Benign

Benign

Phishing

Step02: Input samples



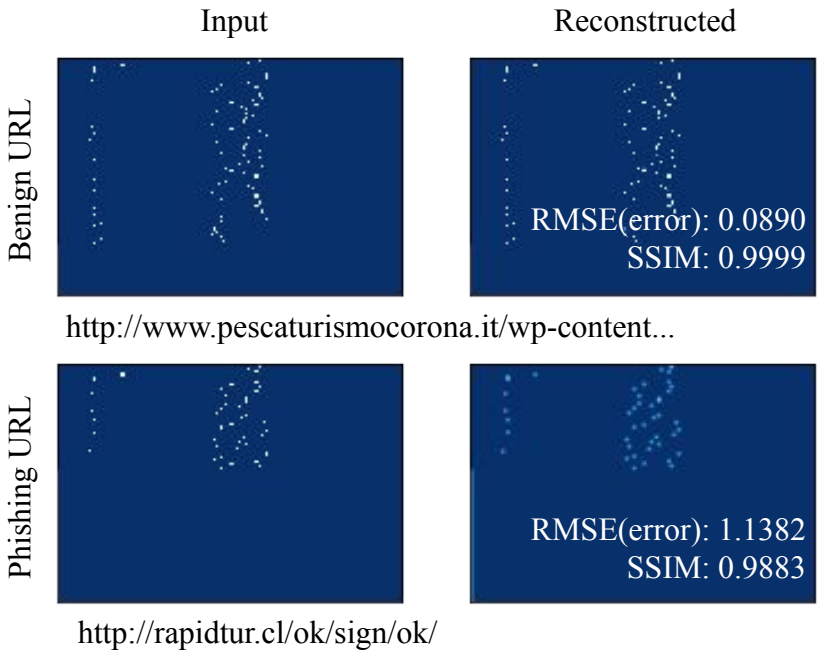
Phishing

Benign

Benign

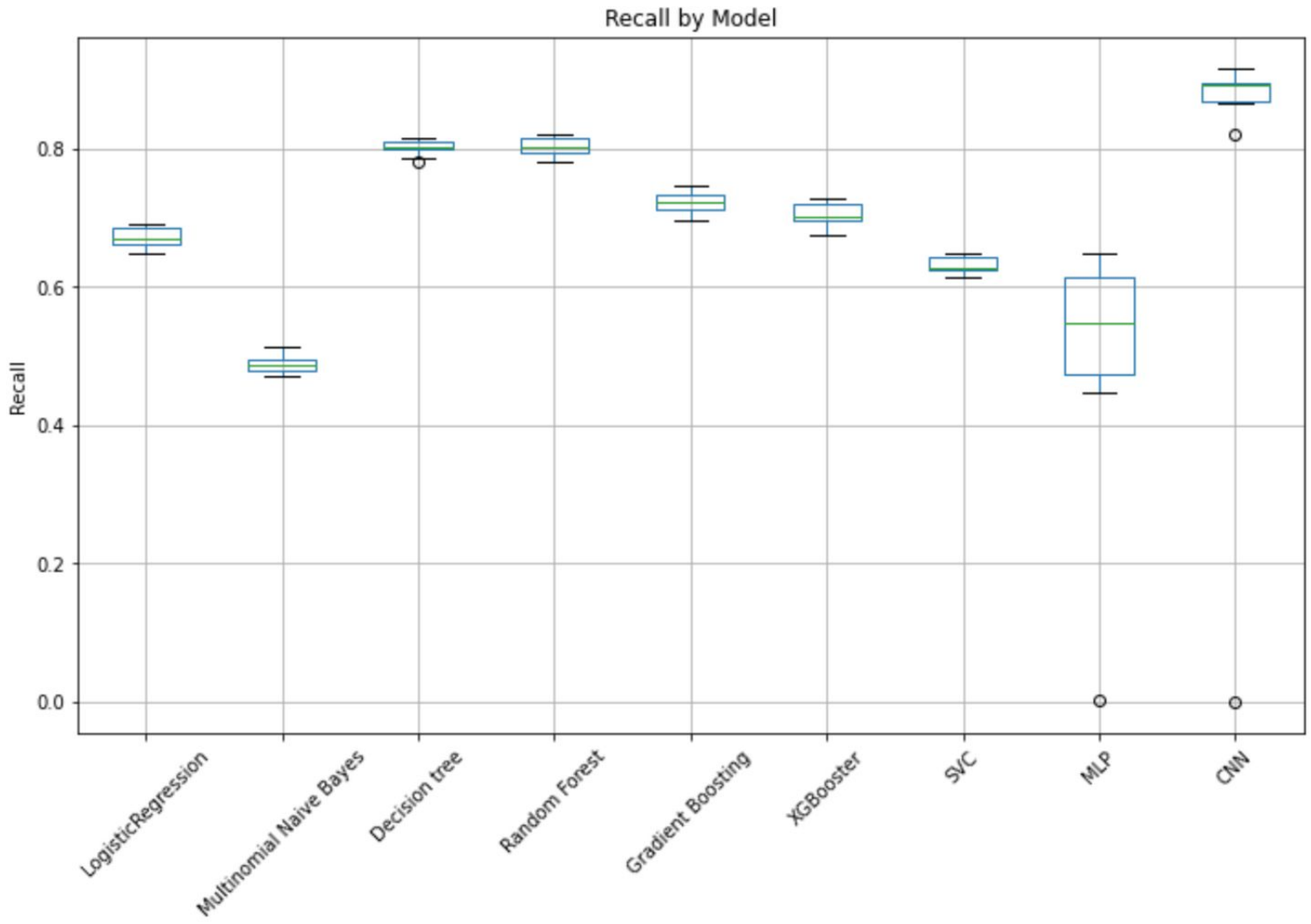
Phishing

정상 vs. 피싱 URL 의 재구축 비교 (1)



	Average RMSE	Average SSIM
Benign URLs	0.5569	0.9995
Phishing URLs	1.1795	0.9986

기계학습 & 딥러닝 모델 Recall Performance 비교



CNN 모델과의 Accuracy / Recall 비교

Ours Model Confusion Matrix

	Positive	Negative
True	13173	373
False	**544**	3910

CNN Model Confusion Matrix

	Positive	Negative
True	13226	320
False	**662**	3792

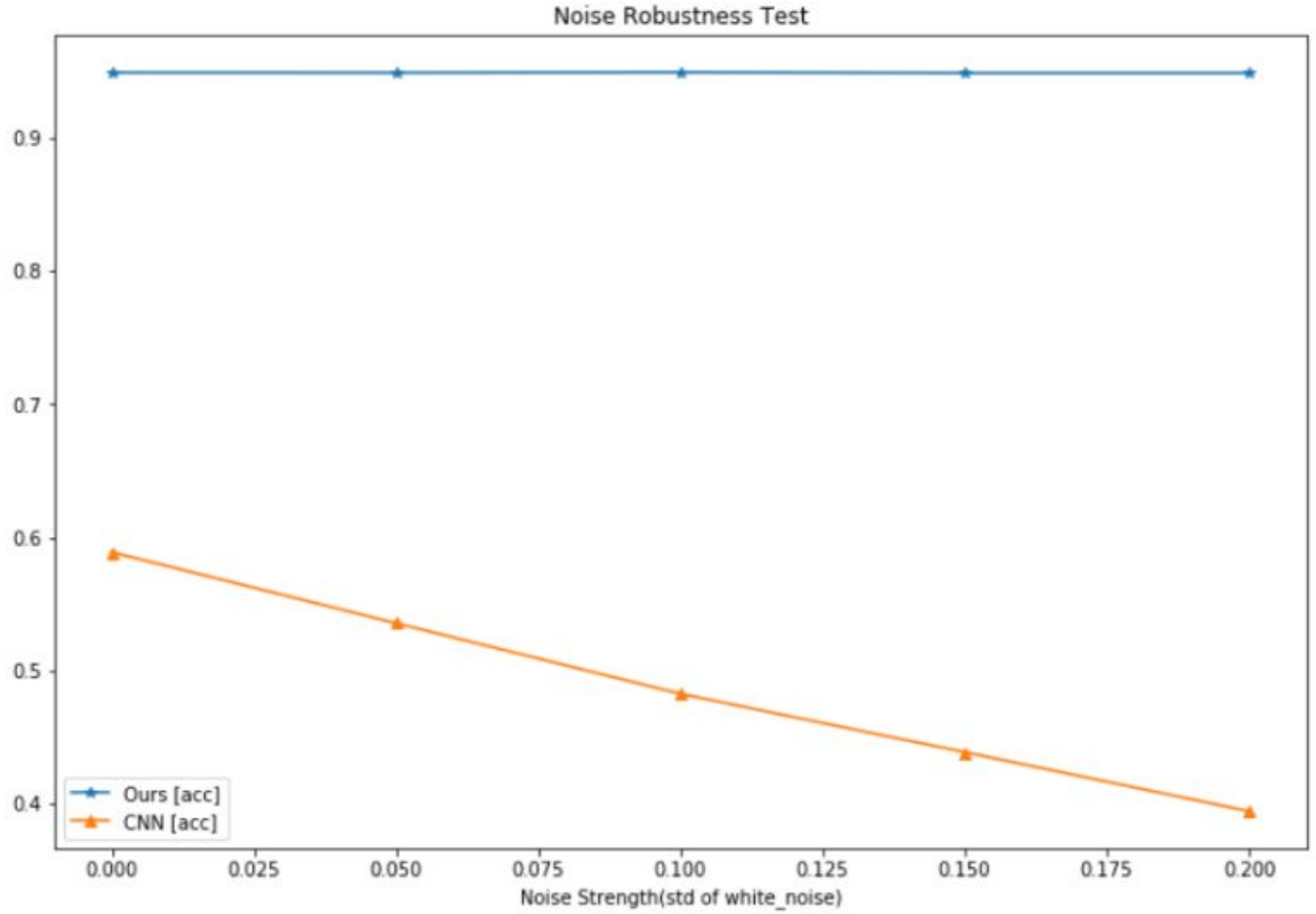
Ours & CNN Classification Report

	precision	recall	f1-score	support
0.0	0.96	0.97	0.97	13546
1.0	0.91	0.88	0.90	4454
accuracy			0.95	18000
macro avg	0.94	0.93	0.93	18000
weighted avg	0.95	0.95	0.95	18000

	precision	recall	f1-score	support
0.0	0.95	0.98	0.96	13546
1.0	0.92	0.85	0.89	4454
accuracy			0.95	18000
macro avg	0.94	0.91	0.92	18000
weighted avg	0.94	0.95	0.94	18000

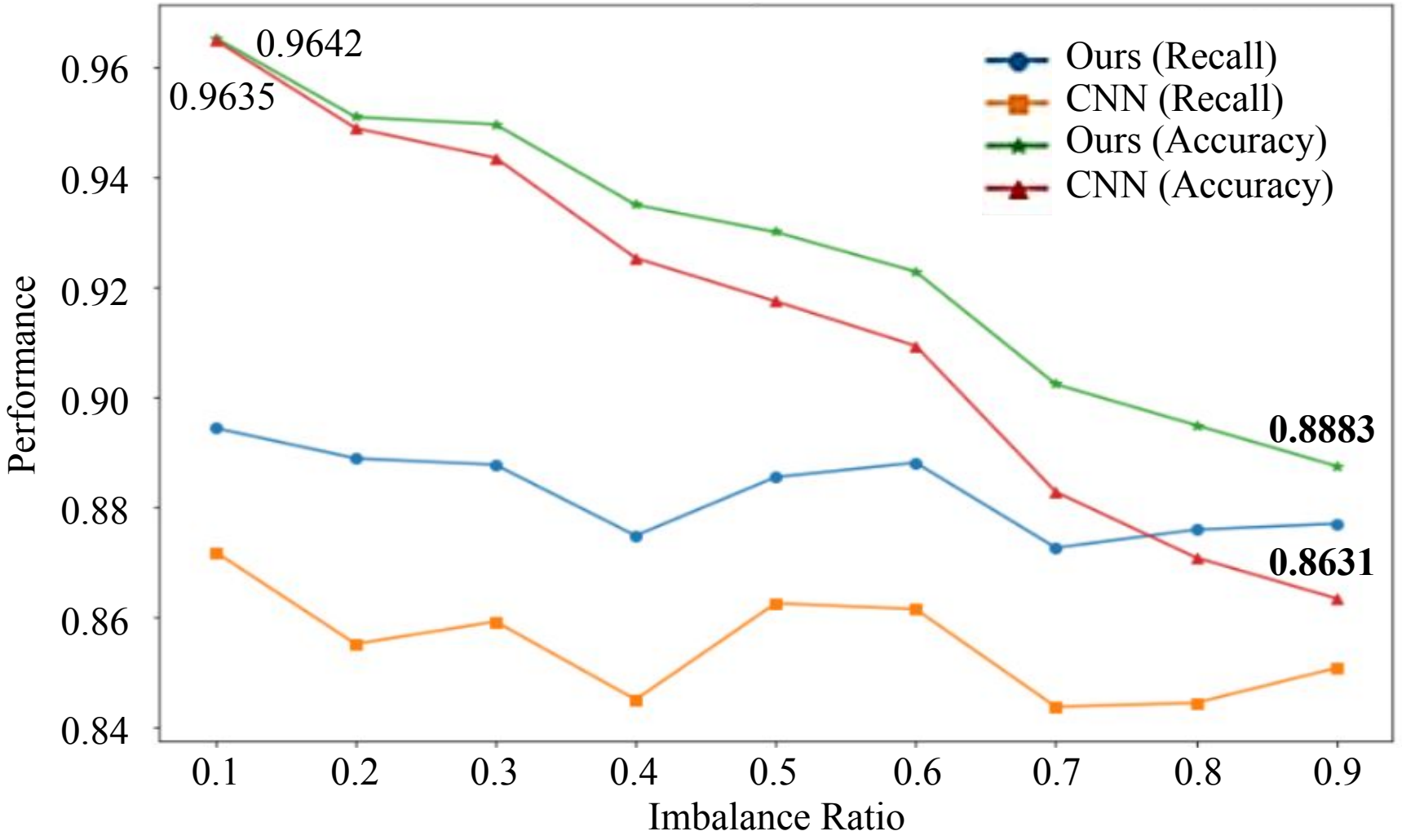
Zero-day Attack에 대한 강건성 평가

- Input URL에 백색 잡음을 추가하는 방법을 통해 강건성 평가 실험 설계
  - 무작위로 생성되는 피싱 URL의 특성을 반영한 실험



Input data imbalance issue 대응 평가

- 이상탐지 모델의 주요 issue 중 하나인 data imbalance 대응 평가 수행
  - Test data 의 피싱 데이터 비율을 증가시키며 performance 측정



결론 및 향후 연구

- Zero-day 특성을 고려한 피싱 URL 의 탐지 방법을 위해 문자 수준의 컨볼루션 오토인코더를 구현하고 정상 URL 의 모형을 학습하였다. 추가로 재구축 오류에 기반한 정상 및 피싱 URL 분류용 컨볼루션 신경망을 제안함 으로서 기존 딥러닝 및 기계학습 알고리즘에 대비하여 최고 정확도를 달성하였고, 10 겹 교차검증하였습니다.

- 제안하는 방법에서는 정상 URL 모형에서 벗어나는 정도를 가늠하기 위한 피싱공격 점수를 가장 직관적인 형태의 유클리디안 거리로 정의하였으나, 특징공간 상에서의 고찰이 필요하다. 또한 오토인코더의 재구축 픽셀공간 상의 잡음에 대한 베이지안 접근 방식의 딥러닝 모형을 고려해볼만 합니다.